AD-A077 828 MILITARY ACADEMY WEST POINT NY OFFICE OF THE DIRECTO--ETC F/G 12/1
USER'S GUIDE TO RESEARCH REPORTS,(U)
NOV 78 T G DAVIDSON
UNCLASSIFIED 79-002 NL

| OF |

AD
A077828

END
DATE
FILMED
1-80
DDC

USER'S GUIDE TO RESEARCH REPORTS

Report Number: 79-002
Project Number: 290
Prepared by: Ted G. Davidson
Typist: Shirley Sabel
NOVEMBER 1978

ABSTRACT

This guide provides operational descriptions of statistical terms, concepts and techniques to assist readers in better understanding research reports which use statistical analysis. The guide is not intended to equip the reader to either conduct statistical analysis or to evaluate the appropriatness of statistical techniques for specific applications. Its sole purpose is to help the reader to more clearly understand the contents of statistical reports.

OFFICE OF
THE DIRECTOR OF INSTITUTIONAL RESEARCH
UNITED STATES MILITARY ACADEMY
WEST POINT, NEW YORK 10996

A

TABLE OF CONTENTS

A. INTRODUCTION:

1. Statistical analysis is an area of mathematics which has evolved from the principles of probability (or relative chance). In general it deals with applying these principles to real-world data and drawing conclusions or inferences on the basis of the relative chance of occurance of various numerical values. The purpose of this guide is to familiarize the layman reader with terms and concepts that are frequently encountered in research reports using statistical analysis. The intent is not to provide precise definitions but rather operational descriptions which will give the reader a grasp of the terms and concepts within the context which they are commonly used. Towards this end the user's guide has been developed based on material contained within the references listed below*, as well as suggestions from researchers in the Office of Institutional Research, United States Military Academy.

2. Most statistical analysis work can be categorized in four general areas: (1) descriptive information of numerical data, (2) estimation of unknown population values from samples, (3) tests of assumptions (or hypotheses), or (4) measures of association between the values which two variables take on. This guide has been organized along these lines. Examples are given throughout where it was deemed useful for the enhancement of reader understanding.

a. Section B of this guide presents descriptions of common statistical terms and establishes a reasonable basis of terminology from which to present the remaining sections.

b. Section C presents statistical measures which are typically used to describe a set of data or the distribution of values of that data.

c. Section D presents statistical tests of hypothesis and significance which appear routinely in statistical work.

d. Section E describes the most commonly encountered measures of association between two variables:

e. Section F discusses some of the more complex techniques of statistical analysis.

3. At the back of this guide is an alphabetized index to aid the reader in locating information about a specific term or technique.

*References:

a. ADKINS, Dorothy C; CONSTRUCTION AND ANALYSIS OF ACHIEVEMENT TESTS; U.S. Government Printing Office; 1947

b. BEYER, William H,; CRC HANDBOOK OF TABLES FOR PROBABILITY AND STATISTICS; The Chemical Rubber Company, 1966

c. ENGLISH, Horace B and ENGLISH, Ava Champney; A COMPREHENSIVE DICTIONARY OF PSYCHOLOGICAL AND PSYCHOANALYTICAL TERMS; David McKay Company Inc; 1958

d. NIE, Norman H, et al: STATISTICAL PACKAGE FOR THE SOCIAL SCIENCES; McGraw-Hill Book Company; 1970

e. SPIEGEL, Murray R; SCHAUM'S OUTLINE SERIES OF STATISTICS; McGraw-Hill Book Company; 1961

B. STATISTICAL TERMS:

1. Webster's New Collegiate Dictionary defines STATISTICS as "1: a branch of mathematics dealing with the collection, analysis, interpretation, and presentation of masses of numerical data." As is common in mathematics, many of the terms are intended to convey precise information about the numerical data. The following descriptions are intended to clarify the meanings of terms which are frequently encountered in research reports using statistical analysis.

2. The following terms describe specific groups associated with "masses of numerical data."

a. The POPULATION is an all inclusive group which may be either finite or infinite in number. The population for a given study is defined within the context of that study and is the group from which samples are taken and to which characteristics of the sample are inferred.

example: If we wish to predict the height of cadets who enter USMA, then the population is all those cadets who have or ever will enter USMA. If we wish to predict the height of applicants to USMA, then the population is all people who have or ever will apply for admission to USMA.

b. A SAMPLE is a subset of the population and implies that every element of the population is not included within the data. It is a finite part of a population whose properties are studied to gain information about the whole population.

example: 10% of all cadets at USMA would be a sample of the population of all cadets at USMA.

c. A RANDOM SAMPLE is a sample selected in such a way that every individual person, object or item in the population has an equal and independent change of being included within the sample.

example: Select 10% of all cadets at USMA by selecting cadets who have a last digit of "7" in their home telephone number.

d. A REPRESENTATIVE SAMPLE is a sample selected in such a way that the sample duplicates the characteristics of the population in all respects that are likely to influence results based on an analysis of the sample.

e. A STRATIFIED SAMPLE is a sample selected by dividing the population into smaller subgroups on the basis of characteristics likely to affect results and taking a number of cases from each subgroup.

example: Select 10% of all cadets at USMA such that the sample includes 10% of each of the four classes.

f. A SYSTEMATIC SAMPLE is a sample selected by taking every k^{th} element in a population.

example: Select 10% of all cadets at USMA by taking every 10^{th} name on a Corps roster.

g. The CONTROL GROUP is a group as closely as possible equivalent to an experimental group and exposed to all the conditions of the investigation except the experimental variable(s) or treatment(s) being studied. Such a group should be representative of the population to which generalization is to be made.

h. The EXPERIMENTAL GROUP is a group exposed to the experimental variable(s) or treatment(s) being studied. It should be representative of the population to which generalization is to be made.

3. The following terms relate information regarding the interpretation of the numerical data.

a. A VARIABLE or VARIATE is anything which can have different numerical values in different individual cases.

b. A CONSTANT is anything which has the same numerical value for each individual case.

c. A RANDOM VARIABLE (or STOCHASTIC VARIABLE) is a variable whose numerical value is the result of chance or random selection.

d. A CONTINUOUS VARIABLE is a variable which can theoretically assume any value between two given values.

example: The measurement of time or distance.

e. A DISCRETE VARIABLE is a variable whose numerical value is confined to a finite or denumerable set of values.

example: The set of integers between one and ten.

f. A DICHOTOMOUS VARIABLE is a discrete variable which may assume exactly two values.

example: Sex as either male or female, responses to true/false questions, or cadets categorized as either upperclassmen or plebes.

g. An INDEPENDENT VARIABLE is a variable whose value is used to predict or estimate the value of another variable. The variable whose value is estimated is called a DEPENDENT or CRITERION VARIABLE. The independent variable is often manipulated by the experimenter to detect relationships with the criterion variable.

h. There are four traditional LEVELS OF MEASUREMENT of data which identify certain ordering and distance properties inherent in the measurement scheme.

(1) NOMINAL-LEVEL MEASUREMENT assigns a name to each distinct category with no assumption of ordering or distance between categories.

example: Sex as male or female, or colors such as red, blue or green.

(2) ORDINAL-LEVEL MEASUREMENT rank-orders each category such that each has a unique position relative to other categories. No assumption of distance between categories is made and, thus, we do not know how much higher or lower one category is than another.

example: Multiple choice responses on a scale of satisfied, neutral or dissatisfied.

(3) INTERVAL-LEVEL MEASUREMENT has the property of rank-ordering and defines the distance between categories in terms of fixed and equal units, but does not have an inherently defined zero point. At this level measures can be added but not multiplied.

example: Temperatures measured in degress fahrenheit or celsius are interval-level measurements.

(4) RATIO-LEVEL MEASUREMENT has all the properties of an interval scale with the additional property that the zero point is inherently defined by the measurement scheme. At this level measures can be added and multiplied.

example: The customary measures of time, distance and weight are ratio-level measurements.

i. An OBJECTIVE MEASURE is a measurement for which the numerical value assigned is unaffected by the opinion or judgement of the scorer. Such a measure is contrasted with a SUBJECTIVE MEASURE for which different scorers may assign different numerical values.

example: Suppose two "scorers" are tasked to count the number of potatoes in a bag. The number of potatoes counted would be the same for both and thus the measure would be objective. If they had been tasked to count the number of "good" potatoes or the number of "big" potatoes in the bag, considerable judgement would have been required and we would not expect the two counts to be equal. Thus, this latter case is an example of a subjective measure.

j. A RAW SCORE is a score as originally obtained, before any transmutation or statistical treatment. Data consisting of such scores are frequently referred to as RAW DATA.

example: The actual time it took a candidate to run the 300-yard shuttle during Physical Aptitude Exam testing is his raw score on that event.

k. A STANDARD SCORE or z-SCORE is a derived score expressing the obtained raw score as a deviation from the mean in standard deviations of the criterion group.

(z = raw score minus mean, divided by standard deviation)

The standard score has a mean of zero and a standard deviation of one.

example: Suppose a candidate receives a raw score of 94 inches on the PAE standing long jump and that the criterion group has a mean of 90 inches and a standard deviation of 8 inches. Then the candidate's standard score on that test would be 0.5 (94-90÷8) which means that his score was one-half a standard deviation higher than the average score for that test.

l. The T-SCALE is a derived score based upon the standard score with a mean of 50 and a standard deviation of 10. Scores of five standard deviations worse or better than the mean are given values of 0 and 100 respectively. Intermediate scores proceed by steps of one for each one-tenth of a standard deviation.

example: A standard score can be converted to the T-scale by multiplying it by ten and adding it from the T-scale mean of 50. Thus, in the above example the standard score of 0.5 would become (0.5)x(10)=5 points which is added to the T-scale mean score of 50. Thus, the standard score of 0.5 on the standing long jump is equivalent to a T-scale score of 55.

m. A STANDARDIZED TEST or STANDARD TEST is one designed to provide a systematic sample of individual performance, administered according to prescribed directions, scored in conformance with definite rules, and interpreted in reference to normative information.

example: The Scholastic Aptitude Tests given to high school seniors by the Educational Testing Service are standardized tests.

4. The following terms relate information regarding the analysis of the numerical data or statistical concepts associated with the analysis.

a. A PARAMETER is a descriptive measure, quantity or value (such as the mean or standard deviation) which is calculated directly from a population or is estimated from a sample and associated with a population.

b. A STATISTIC is a value or number that describes a series of quantitative observations or measures, or a value calculated from a sample that is supposed to describe the population from which the sample is drawn.

(1) DESCRIPTIVE STATISTICS are those statistics used only for the purpose of describing the sample from which they are derived.

(2) INFERENTIAL STATISTICS are those used to infer characteristics of the population from which the sample is drawn.

c. A FREQUENCY DISTRIBUTION is a systematic grouping of data into categories according to the frequency of occurence of each data item in each category. A COMMULATIVE FREQUENCY DISTRIBUTION is a tabulation showing how many data items fall at or below each of the successive values arranged in order of magnitude, and a graph of which forms an OGIVE.

(1) A DISCRETE DISTRIBUTION is a frequency distribution for which the numerical values of the data are confined to a finite or denumerable set of values.

(2) A CONTINUOUS DISTRIBUTION is a frequency distribution for which the numerical values of the data can theoretically assume any values between two given values.

(3) The NORMAL DISTRIBUTION or GAUSSIAN DISTRIBUTION is a bell-shaped continuous distribution which is widely used in statistics because it often approximates observed phenomenon.

(4) DISTRIBUTION-FREE or NONPARAMETRIC methods are methods of analyzing data that make no assumptions concerning the shape of the true distribution.

d. A PERCENTILE is the point in a frequency distribution below which that percentage of cases fall. Thus, 62 percent of the cases fall below the sixty-second percentile. A DECILE refers to every tenth percentile. A QUINTILE to every twentieth percentile, and a QUARTILE to every twenty-fifth percentile.

e. A NULL HYPOTHESIS is an assumption by the researcher which he seeks to disprove. The null hypothesis is the logical complement of the RESEARCH HYPOTHESIS which one seeks to prove. It is possible to disprove the null hypothesis, whereas it is impossible to prove the original research hypothesis directly. Hence, a research design commonly calls for a test to see whether the null hypothesis can be disproved. If so then the research hypothesis is confirmed. Failure to disprove the null hypothesis does not permit any inference about the research hypothesis.

example: Suppose a researcher subjects an experimental group to a treatment and wishes to show a difference between the means of the control group and the experimental group. His research hypothesis would be that the groups' means are different, but his null hypothesis would be that the difference in means is no larger than could be expected by chance alone. If the difference of the means is statistically significant, the researcher can conclude that there is very little chance that the null hypothesis is true and, therefore, the research hypothesis must be true.

f. STATISTICAL SIGNIFICANCE is the probability that the value actually obtained would occur by chance alone if the null hypothesis were true. A low probability can be attributed to something other than chance. Statistical significance does not imply practical significance.

g. The POWER of a statistical test is the probability of rejecting the null hypothesis when some other hypothesis is true (probability of avoiding a type II error).

h. TYPE I ERROR refers to rejecting a null hypothesis which is in reality true. TYPE II ERROR refers to accepting a null hypothesis which is in reality false.

i. VALIDITY refers to the property that the obtained scores correctly measure the variable they are supposed to measure.

j. RELIABILITY is the measure of consistency between results of repeated administrations of the same measuring device to the same individuals and is usually estimated in terms of the COEFFICIENT OF RELIABILITY or of the standard error of measurement.

k. COEFFICIENT OF RELIABILITY is an estimate of the correlation between repeated administrations of the same measuring device to the same individuals without distrubances by such factors as memory, practice, boredom, etc.

l. CORRELATION is a measure of association of change between two variables, normally expressed as a CORRELATION COEFFICIENT. Positive correlation implies that the two variables tend to increase or decrease together. Negative correlation implies that an increase in one variable tends to accompany a decrease in the other variable.

m. A ROBUST procedure is one which is relatively insensitive to departures from its assumptions.

n. The REGRESSION EFFECT or REGRESSION TOWARD THE MEAN is the tendency for a group, selected as being any given amount above or below the mean on one test, to be closer to the mean on a second test. It results from the fact that a certain number of individuals are found in the specified range partly by reason of errors of measurement which will not affect them on another test. The regression effect does not affect the range, mean, or standard deviation of the entire distribution but only that of the selected subgroup.

example: Suppose that we retested all candidates who were between one and two standard deviations below the mean on the PAE test and found that 10% of them scored higher than one standard deviation below the mean. This result does not necessarily mean that their mean performance level has increased but rather is a predictable regression effect.

o. CROSS-VALIDATION or VALIDATION is the application of a procedure found to work with one sample by trying it out on a second sample of the population in question to see if the results are similar.

C. STATISTICAL MEASURES OF A DISTRIBUTION:

1. Statistical measures are used to convey useful attributes of masses of data. The following measures are the most common for describing the central tendency, the dispersion and the shape of the frequency distribution associated with the data.

2. The following are MEASURES OF CENTRAL TENDENCY which tend to lie centrally within a set of data arranged according to magnitude.

a. The ARITHMETIC MEAN or briefly the MEAN (commonly called "average") of a set of N numbers X_1, X_2, ---, X_N is denoted by $\overline{X}$ (read "X bar") and is defined as

$$\overline{X} = \frac{X_1+X_2 + X_3 + --- + X_N}{N} = \sum \frac{X}{N}$$

when it is necessary to distinguish the mean of a population from the mean of a sample drawn from this population, μ(mu) is customarily used to represent the population mean. In statistics the mean is frequently referred to as the EXPECTED VALUE of a distribution.

b. The MEDIAN is that number which separates the smallest 50% of the set from the largest 50%.

c. The MODE of a set of numbers is that value which occurs with the greatest frequency. The mode may not exist, and even if it does exist it may not be unique. A distribution having only one mode is called UNIMODAL.

3. The following MEASURES OF DISPERSION indicate the degree to which the numerical data tend to spread about a typical central value.

a. The RANGE of a set of numbers is the difference between the largest and the smallest numbers in the set. The range is a crude measure of the variability of the set of numbers.

b. The INTERQUARTILE RANGE is the difference between the first quartile and the third quartile of the set of numbers and contains the middle 50 percent of the cases in the distribution.

c. The VARIANCE of a set of N numbers $X_1, X_2, ---, X_N$ is denoted by S^2 and is defined by

$$S^2 = \frac{(X_1 - \overline{X})^2+(X_2 - \overline{X})^2+ --- + (X_N - \overline{X})^2}{N}$$

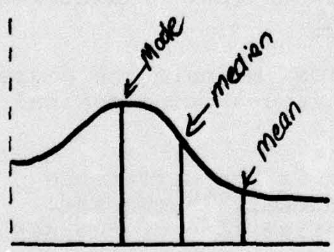
where $\overline{X}$ is the mean of the set of numbers. When it is necessary to distinguish the variance of a population from the variance of a sample drawn from this population, S^2 is customarily used to represent sample variance and σ^2 (sigma-squared) to represent population variance.

d. The STANDARD DEVIATION of a set of numbers is defined as the square root of the variance. For a normal (bell-shaped) distribution, 68.3% of the cases fall within plus-or-minus one standard deviation of the mean, 95.5% fall within plus-or-minus two standard deviations of the mean and 99.7% fall within plus-or-minus three standard deviations of the mean.

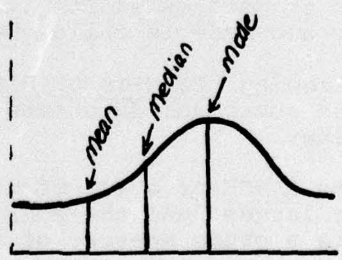
e. The STANDARD ERROR is the estimated standard deviation of the values of a statistic (such as the sample mean) that would be obtained if the measurement were repeated over and over again.

4. The following MEASURES OF SHAPE indicated the degree of symmetry and peakedness associated with the frequency distribution of the numerical scores.

a. SKEWNESS is a measure of the degree of departure from symmetry of a unimodal distribution. If the distribution has a longer tail to the right of the central maximum than to the left, the distribution is said to be SKEWED TO THE RIGHT or to have POSITIVE SKEWNESS. If the reverse is true it is said to be SKEWED TO THE LEFT or to have NEGATIVE SKEWNESS. The measure of skewness will take on a value of zero when the distribution is a completely symmetric curve. The figures below show the relative positions of the mean, median and mode for distributions which are skewed to the right and left respectively. For symmetrical curves the mean, median and mode coincide.



skewed right



skewed left

b. KURTOSIS is a measure of the degree of peakedness of a distribution. A normal distribution will have a kurtosis of zero. If the distribution is more peaked than the normal distribution the kurtosis is POSITIVE. If it is flatter the kurtosis is NEGATIVE.

D. TESTS OF HYPOTHESIS AND SIGNIFICANCE:

1. In statistical analysis, the objective is sometimes merely to estimate unknown population parameters. But more often the ultimate purpose will involve some use of the estimate to make decisions about populations on the basis of sample information. Such decisions are called STATISTICAL DECISIONS. In attempting to reach decisions it is customary to make assumptions or NULL HYPOTHESES (denoted H_o) about the population involved. If on the supposition that the null hypothesis is true we find that results observed in a random sample differ markedly from those expected on the basis of pure chance, we say that the observed differences are STATISTICALLY SIGNIFICANT and we are inclined to reject the null hypothesis. Any hypothesis which differs from the null hypothesis is called an ALTERNATE HYPOTHESIS and is customarily denoted by H_1. Procedures which enable us to decide whether to accept or reject hypotheses or to determine whether observed samples differ significantly from expected results are called TESTS OF HYPOTHESES, TESTS OF SIGNIFICANCE, or RULES OF DECISION. The statistic which is compared to the expected results is called the TEST STATISTIC.

2. In testing a given hypothesis, the maximum probability with which we are willing to risk rejecting the null hypothesis when it is true is called the LEVEL OF SIGNIFICANCE of the test and is often denoted by α (alpha). In practice a level of significance of .05 or .01 is customary, although other values can be used.

3. The t-TEST is a test of the difference between two sample means. It is based upon the "Student's t" distribution from which critical values are derived for the desired level of significance. If the observed t statistic exceeds the critical value it is concluded that the difference between the means is too large to be attributed to chance alone, and that the two samples are representative of different parent populations.

4. The CHI-SQUARE (χ^2) TEST is a test of the difference between observed frequencies and expected frequencies computed on the basis of a null hypothesis. It is based upon critical values derived from the χ^2 distribution which the χ^2 statistic very closely approximates if expected frequencies are at least equal to 5. If the χ^2 statistic exceeds the critical value for the applicable level of significance, it is concluded that the difference between the observed frequencies and expected frequencies is too large to be attributed to chance alone and the null hypothesis is rejected. A significant χ^2 statistic does not indicate how strongly variables are related nor which of a set of variables is responsible for the statistically significant result.

5. The F-TEST is a testing procedure based upon the "Snedecor's F" distribution which is the ratio of two independent χ^2 distributed variables. This test is widely used in statistics because of the numerous ways in which null hypotheses can result in two independent χ^2 distributed statistics. The

critical value for the desired level of significance is derived from the known F distribution. If the observed F value exceeds this level it is concluded that the difference is too large to be attributed to chance alone and the null hypothesis is rejected.

6. NONPARAMETRIC TESTS are methods of analyzing data and making statistical decisions which do not make any assumptions concerning the true distribution of the variable. In contrast, the derivations of the "Student's t", the χ^2 and the "Snedecor's F" distributions all rely upon the assumption of an underlying distribution which is normal (or near normal). Consequently, when the true distribution is not normal these test statistic distributions only approximate the distributions from which the critical values are derived.

E. MEASURES OF ASSOCIATION:

1. In contrast to tests of significance, which are probability statements regarding the actual existence of relationships within the population, measures-of-association indicate how strongly two variables are related to each other. CORRELATION is the tendency of two variables to vary together, and a CORRELATION COEFFICIENT is a single number which summarizes the strength of this relationship. Unless otherwise implied by the context within which it is used, the term correlation coefficient customarily implies the Pearson product-moment correlation coefficient.

2. The PEARSON PRODUCT-MOMENT CORRELATION COEFFICIENT (r) is the most common method of determining a relationship between two variables. It is applicable when two variables are both measured at the interval or ratio level, and takes on values between plus one and minus one. Positive values imply the variables increase or decrease concomitantly (direct correlation), negative values imply that an increase in one variable tends to accompany a decrease in the other (inverse correlation), and zero implies that the two variables vary independently. A correlation coefficient of plus one minus one indicates a perfect relationship, zero indicates no linear relationship, and intermediate values imply relationships which are proportionately less perfect. The square of Person's r (denoted by r^2) is a more easily interpreted measure of association when our concern is with the strength of relationship rather than direction of relationship. r^2 has a range of 0 to 1 and is a measure of the proportion of variance in one variable "explained" by the other (often referred to as SHARED VARIANCE).

3. The POINT BISERIAL CORRELATION COEFFICIENT (r_p or r_{pb}) is the product-moment correlation between a continuous variable and another variable represented by a dichotomy.

4. The CORRELATION RATIO (ETA) is a measure of association used when the independent (or predictor) variable is nominal level and the dependent (or criterion) variable is interval or ratio level. It is basically an indication of how dissimilar the means on the dependent variable are within the categories of the independent variable. ETA-SQUARED has an
intuitive
intuitive interpretation as the proportion of variance in the dependent variable explained (or accounted for) by the independent variable. Eta-square will always be larger than r^2 and is a useful measure of the extent of correlation when variable relationships are curvilinear. The amount by which eta-square exceeds r^2 is a reflection of the appropriatness of assuming a linear relationship between the variables.

5. SPEARMAN'S RHO (r_s) and KENDALL'S TAU are nonparametric correlations based on the ranking of two ordinal level variables. They presume a large number of categories (or ranks) and a small number of ties in ranking. Both

16

coefficients range from plus one to minus one. The chief difference between Spearman's r_s and Kendall's tau is that the Kendall coefficient is somewhat more meaningful when the data contain a large number of tied ranks.

F. TECHNIQUES OF STATISTICAL ANALYSIS:

1. Some of the more frequently encountered techniques of statistical analysis are described below. In general they are methods for summarizing and conveying the essence of relationships which exist within the data and which may be inferred to exist within the parent populations which the data samples represent. The examples given are for illustration only and are not based on research actually conducted.

2. CONTENT ANALYSIS consists of discovering and listing according to a systematic plan the ideas and feelings represented in free-response statements. An objective tabulation of the frequency with which certain elements occur is usually included.

3. A CROSSTABULATION (or CONTINGENCY TABLE) is a joint frequency distribution of cases according to two or more classificatory variables. These joint frequency distributions can be analyzed statistically by tests of significance to determine whether or not the variables are independent and can be summarized by measures of association which describe the degree to which values of one variable change with those of another.

example: Suppose we conducted a survey of 200 people to gather data about color preference and age. The following crosstabulation shows the frequency with which persons of various ages preferred the various colors.

		COLOR PREFERRED			
		RED	GREEN	BLUE	TOTAL
AGE	10-19	30	5	10	45
	20-29	10	20	30	60
	30-39	20	10	25	55
	over 39	5	15	20	40
	total	65	50	85	200

4. An EXPECTANCY TABLE is one which gives the probabilities of various scores on a criterion variable as a function of scores on a predictor variable.

example: Suppose that the following expectancy table represents the relationship between the number of children in a family (the predictor variable) and the number of cars owned by the family (the criterion variable).

		NUMBER OF CARS					
		0	1	2	3	over 3	total
NUMBER OF CHILDREN	0	.08	.70	.22	.00	.00	1.00
	1	.06	.64	.28	.02	.00	1.00
	2	.02	.61	.29	.06	.02	1.00
	over 2	.00	.52	.32	.12	.04	1.00

From the table we see that the number of cars owned by a family appears to be related to the number of children in the family. For a family with one child, the probability of owning exactly 2 cars is .28. The same probability for a family with more than 2 children is .32, and for no children it is .22.

5. A SCATTERGRAM is a graph or plot of data points based on two variables, where one variable defines the horizontal axis and the other defines the vertical axis.

6. REGRESSION ANALYSIS is a method for predicting a criterion variable from scores on one or more predictor variables. The prediction is based upon the correlation of each of the predictor variables with the criterion variable and on their inter-correlations. The equation for predicting the criterion variable (the REGRESSION EQUATION) is the particular weighted sum of the predictor variables which correlates highest with the criterion variable. This correlation is called the COEFFICIENT OF MULTIPLE CORRELATION (R) and its square (R^2) is the proportion of variance in the criterion variable which is explained or accounted for by the predictor variables (SHARED VARIANCE).

example: We wish to predict the highway cruising speed (HCS) for cars of various types. We suspect that the speed is at least partially determined by (1) car weight (WT), (2) engine horsepower (HP), (3) number of miles the car has been driven (NM) and (4) the age of the driver (AGE). Therefore, we collect data on a sample of cars driving down the highway and calculate the following regression equation for predicting cruising speed.

$$HCS_P = .00625\ (WT) + .135\ (HP) - .00025\ (NM) + 4.5$$

$$R = .70 \qquad R^2 = .49$$

This prediction has a higher correlation with actual highway cruising speed than does any other weighted sum of car weight, horsepower, mileage of car and age of driver and accounts for 49% of the variance of cruising speeds. The regression equation shows that, all else being equal, an additional 1000 pounds of car weight would be expected to be accompanied by a 6.25 MPH increase in highway cruising speed. Similarly an increase of 10 horsepower in engine size would be accompanied by a 1.35 MPH increase in predicted speed and an increase of 10,000 miles in car mileage would be accompanied by a decrease of 2.5 MPH in predicted cruising speed. The 4.5 constant is a scaling factor which adjusts the mean of the prediction to the mean of the observed cruising speeds. The absence of AGE in the regression equation indicates that the age of the driver contributes no additional information regarding highway cruising speed. Thus the best prediction of highway cruising speed for a car which weights 4000 pounds, has a 300 horsepower engine and has 40,000 miles on it is:

$$.00625\ (4000) + .135\ (300) - .00025\ (40000) + 4.5$$

which equals 60 miles per hour.

7. ANALYSIS OF VARIANCE (ANOVA) is a method for testing the equality of the means of 3 or more groups. Usually the groups are subjected to different experimental treatments and the ANOVA tests whether the means produced by the treatments are equal. If they are not, it is concluded that the treatment affected the criterion. If the analysis concerns the effects of a single treatment it is termed ONE-WAY ANALYSIS OF VARIANCE. If it concerns n different treatments it is referred to as n-WAY ANALYSIS OF VARIANCE.

example: We wish to determine whether the color of office walls have any effect on the typing speed of typists. We have decided to test red walls, green walls, blue walls and white walls. Therefore we selected four similar offices, painted each with one of the treatment colors and collected data on the typing speeds in each office. We did a one-way analysis of variance on the data and concluded that the differences of the mean typing speed of the four offices are statistically significant. This result indicates that the observed differences in mean typing speeds are too large to be attributed to chance alone and that the color of the office wall does affect the speed of the typist. The ANOVA itself does not tell which color is better nor how much change in typing speed is associated with the colors tested. However, other statistical techniques are available to investigate these concerns.

8. DISCRIMINANT ANALYSIS is a method for distinguishing between two or more groups of cases by using a weighted sum of variables. One or more such weighted sums (DISCRIMINANT FUNCTIONS) are formed in such a way as to maximize the separation of the groups. Once the discriminant functions have been derived, statistical tests can be conducted to measure the success with which the variables actually distinguish among the groups. After satisfactory discrimination for cases with known group memberships has been established, a set of CLASSIFICATION FUNCTIONS (one for each group) can be derived which will give the probability of group membership for new cases with unknown memberships.

example: We wish to distinguish between three groups of people on the basis of their typical voting pattern: (1) those who typically vote Democratic, (2) those who typically vote Republican, and (3) those who typically vote Independent. We suspect that we can distinguish between these voting patterns on the basis of the following five discriminating variables: (1) their annual income (INC), (2) their age (AGE), (3) the size of the community in which they live (SIZE), (4) the distance from their home to the nearest city with a population of 500,000 or more (DIST), and (5) the length of their hair (HAIR). Therefore, we collect data on a sample of voters and conduct a discriminant analysis

20

which yields the following two statistically significant discriminant functions (D):

DF1 = .039(INC) - .023(DIST) - .231(HAIR) + 0.986

DF2 = .064(INC) - .013(SIZE) + .133(HAIR) - 1.677

We conclude that we can distinguish between groups on the basis of these variables and that the difference between groups can be characterized as two uncorrelated underlying characteristics (one for each discriminant function). We also note that age did not contribute to discrimination among groups since it did not appear in either discriminant function, and that income and length of hair contribute to both underlying characteristics since they appear in both discriminant functions. We validate the results by applying the two discriminant functions to a second sample of data and find that we correctly identify the voting pattern in 85% of the cases. Thus, the success with which the variables actually distinguish among the voting patterns is substantiated and the results of the analysis can be used to predict the voting pattern of people whose voting pattern is unknown.

9. FACTOR ANALYSIS is a general procedure for identifying and defining dimensional space among a group of variables. Its major uses are to identify a smaller number of valid dimensions (FACTORS) contained in a set of variables, and to determine the degree to which given variables are part of a common underlying phenomenon. Typically, factor analysis is applied to a large number of variables to derive one or more weighted sums of the variables (FACTORS) which account for a substantial portion of the total variance. This smaller number of factors contains most of the information in the large number of variables and is customarily used in lieu of the large number of variables in subsequent analyses. Frequently individual factors can be interpreted as representing identifiable characteristics or dimensions about which the set of variables provides information.

example: Suppose that we have been tasked to evaluate the employee performance appraisal system of a large company. In the initial phase of the investigation, we discover that each ratee is graded on 7 job duties. We strongly suspect that 7 separate scores are excessive for describing an employee's performance during a rating period and wish to reduce the number to one that is somewhat more manageable and interpretable without excessive loss of information contained within the scores. The performance appraisal data is already computerized (7 scores for each employee which we designate VAR1 through VAR7 respectively) so we decide to conduct a factor analysis of this data to investigate the possibility of reducing the required number of appraisals and to derive combinations of appraisal scores (called factors) which can be used as variables in further evaluation of the appraisal system. The factor analysis results in the following three initial factors which account for 78% of the variance in the appraisal scores:

FACT1 = .68(VAR1)+.51(VAR2)+.37(VAR3)+.81(VAR4)-
.23(VAR5)+.82(VAR6)-.26(VAR7)

FACT2 = .38(VAR1)+.63(VAR2)-.43(VAR3)+.42(VAR4)+
.65(VAR5)-.43(VAR6)+.53(VAR7)

FACT3 = .41(VAR1)+.57(VAR2)+.48(VAR3)-.31(VAR4)-
.39(VAR5)+.10(VAR6)+.61(VAR7)

Therefore, we conclude that most of the information contained in the 7 appraisal scores is also contained in these three factor scores. The three factors can be interpreted as representing three uncorrelated underlying characteristics (or dimensions) of employee performance and may be visualized as the three axes of a 3-dimensional coordinate system. The factors are not unique because the coordinate system can be rotated such that the factors take on different weighted sums of the 7 variables but, as a set, retain the same information regarding employee performance. We conduct such a rotation of the axes to derive factors which are simplified and more interpretable. The rotation results in the following three terminal factors:

FACT1 = .86(VAR1)+.12(VAR2)+.03(VAR3)+.69(VAR4)-
.05(VAR5)+.92(VAR6)-.07(VAR7)

FACT2 = .18(VAR1)+.97(VAR2)-.06(VAR3)+.23(VAR4)+
.77(VAR5)+.09(VAR6)-.17(VAR7)

FACT3 = .03(VAR1)-.16(VAR2)+.74(VAR3)+.63(VAR4)+
.18(VAR5)-.11(VAR6)+.83(VAR7)

We note that each variable now loads either very high or very low on each factor. FACT1 is composed predominantly of VAR1, VAR4 and VAR6 (factor loadings of .86, .69 and .92 respectively). Similarly, FACT2 is composed predominantly of VAR2 and VAR5 while FACT3 is predominantly VAR3, VAR4 and VAR7. Therefore, we simplify the factors by retaining only the dominant variables. This results in the following three simplified factors which should closely approximate the three terminal factors:

FACT1 = .86(VAR1) + .69(VAR4)+.92(VAR6)

FACT2 = .97(VAR2) + .77(VAR5)

FACT3 = .74(VAR3) + .63(VAR4) + .83(VAR7)

To ensure that the approximation is satisfactory we correlate each of the simplified factors with its corresponding terminal factor and find correlations of .99, .96 and .98 respectively. The square of these correlations (.98, .92 and .96) show that over 90% of the variance (or information) contained in the terminal factors is retained by the simplified factors. We can now conclude that the employee performance data being collected can be typified as three uncorrelated underlying dimensions which are adequately represented by the three simplified factors, and we can use those three factors instead of the cumbersome 7 variables which we started out with.

10. CANONICAL CORRELATION ANALYSIS is in some respects a generalization of multiple regression. It takes two sets of variables, each of which can be given theoretical meaning as a set, and derives a weighted sum (called a CANONICAL VARIATE) from each of the sets in such a way that the correlation between the two combinations is maximized. Several such pairs of weighted sums may be derived, each successively accounting for the maximum amount of relationship remaining (RESIDUAL VARIANCE) between the two sets of variables. Since successive pairs of canonical variates account for residual variance, the pairs are uncorrelated with one another. The most important information produced by canonical correlation analysis are the CANONICAL VARIATES (weighted sums of variables) and the CANONICAL CORRELATIONS (correlations between corresponding pairs of canonical variates). The square of a canonical correlation (called an EIGENVALUE) represents the proportion of variance in one canonical variate that is accounted for by its paired canonical variate.

example: We are interested in identifying employees who can be expected to perform exceptionally well if promoted into managerial positions. We have analyzed the positions and concluded that performance consists predominantly of four types of activities: (1) motivation of subordinates (MOT), (2) personal direction of project and program teams (DIR), (3) planning for future contingencies and projects (PLA) and (4) administration of routine correspondence and policies (ADM). We are convinced from past experience that aptitude for exceptional managerial performance is embedded primarily in four attributes: (1) the highest education level achieved (ED), (2) the age of the employee (AGE), (3) the innate intelligence of the employee (IQ) and (4) the employee's previous performance history (PH). Therefore, we gathered data on these eight variables and conducted a canonical correlation analysis using the four performance variables (MOT, DIR, PLA and ADM) as one set of variables and the four aptitude variables (ED, AGE, IQ and PH) as the other set of variables. The analysis resulted in the following three statistically significant pairs of canonical variates:

PAIR1: P=.03(MOT)+.51(DIR)-.18(PLA)+.72(ADM)
A=.12(ED)+.43(AGE)-.17(IQ)+.67(PH)

R=.68 R^2=.46

PAIR2: P=.68(MOT)+.12(DIR)+.62(PLA)-.13(ADM)
A=.62(ED)-.13(AGE)+.57(IQ)+.03(PH)

R=.48 R^2=.23

PAIR3: P=.43(MOT)-.57(DIR)+.27(PLA)-.35(ADM)
A=.29(ED)-.43(AGE)+.34(IQ)-.53(PH)

R=.11 R^2=.01

We note that the first pair of canonical variates is predominately a relationship between the performance variables DIR and ADM and the aptitude variables AGE and PH. Since this is the first pair we conclude that the managerial performance activities

of directing and administration are the most predictable aspects of the managerial activities and that they relate predominately to the age and previous performance history of the incumbent. Similarly, the second pair of canonical variates can be interpreted as representing a dimension of managerial performance which is uncorrelated to the first pair and which is primarily a relationship between the motivating and planning performance of the manager and his education and innate intelligence. The third canonical pair appears to be a conglomeration in which none of the variables clearly dominate. Although it is statistically significant the canonical correlation ($R=.11$) is quite low and the shared variance ($R^2=.01$) indicates that the variates share only 1% of their variance. Consequently, we decide to disregard the third canonical pair and accept the first two pairs as representative of the valid relationship which exists between the four performance variables and the four aptitude variables.

INDEX

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: 14 79-ØØ2

2. GOVT ACCESSION NO.

3. RECIPIENT'S CATALOG NUMBER

4. TITLE (and Subtitle): 6 USER'S GUIDE TO RESEARCH REPORTS,

5. TYPE OF REPORT & PERIOD COVERED

6. PERFORMING ORG. REPORT NUMBER

7. AUTHOR(s): 10 [redacted] Ted G. Davidson

8. CONTRACT OR GRANT NUMBER(s): 12 28

9. PERFORMING ORGANIZATION NAME AND ADDRESS

Office of the Director of Institutional Research
United States Military Academy
West Point, New York 10996

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS

11. CONTROLLING OFFICE NAME AND ADDRESS

Same as #9 above.

12. REPORT DATE: 11 Nov [redacted] 78

13. NUMBER OF PAGES: 25+i

14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)

United States Military Academy
West Point, New York 10996

15. SECURITY CLASS. (of this report): Unclassified

15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Reproduction of this document in whole or in part must have prior approval of the Superintendent, United States Military Academy, West Point, New York.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

This guide provides operational descriptions of statistical terms, concepts and techniques to assist readers in better understanding research reports which use statistical analysis. The guide is not intended to equip the reader to either conduct statistical analysis or to evaluate the appropriatness of statistical techniques for specific applications. Its sole purpose is to help the reader to more clearly understand the contents of statistical reports.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

406 247